Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy.

Devillé, W.L.J.M.; Bezemer, P.D.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2000

DOI (link to publisher)
10.1016/S0895-4356(99)00144-4
S0895-4356(99)00144-4 [pii]

document version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Clinical Epidemiology 53 (2000) 65–69

Journal of Clinical Epidemiology

Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy

W.L.J.M. Devillé[a,*], P.D. Bezemer[a], L.M. Bouter[b]

[a]*Department of Epidemiology and Biostatistics, Medical Faculty, Vrije Universiteit Amsterdam, Van der Boechorststraat 7, NL-1081 BT Amsterdam, The Netherlands;*

[b]*Institute for Research in Extramural Medicine, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*

Received 20 October 1998; received in revised form 1 July 1999; accepted 1 July 1999

Abstract

Search strategies for articles reporting on diagnostic test evaluations have been subjected to less research than those in the domain of clinical trials. We set out to develop an optimal search strategy for publications on diagnostic test evaluations in general, that could be added to keywords describing the specific diagnostic test at issue. Nine Family Medicine journals were searched from 1992 through 1995 for primary publications on diagnostic test evaluation by hand searching and a Medline search strategy published earlier. Additionally, new search strategies have been developed with stepwise logistic regression, using Mesh terms and free text words related to diagnosis and test evaluation as independent variables. Hand searching identified 75 primary publications on diagnostic test evaluation from a total of 2467 primary publications. The previously published search strategy had a sensitivity of 73%, a specificity of 94%, and a positive predictive value of 29%. The most accurate new search strategy had a sensitivity of 80.0% (60/75; 95% CI: 71.0–89.1), a specificity of 97.3% (2327/2392; 95% CI; 96.6–97.9%), a positive predictive value of 48% (95% CI: 40–56) and diagnostic odds ratio of 149. All four new strategies used the Mesh term "sensitivity and specificity" (exploded with the Mesh terms "predictive value" and "ROC")and cumulatively added the text words "specificity," "false negative," "accuracy," and "screening." The search strategy using the Mesh term "sensitivity and specificity" (exploded) and the text words "specificity," "false negative," and "accuracy" has both higher sensitivity and specificity than the previously published strategy. The increase in specificity in three strategies reduces the absolute number of false-positive articles that have to be screened by 50–75%, compared to the number of false positives in the earlier strategy.

Keywords: Diagnosis; Sensitivity and specificity; Medline; Family medicine; Logistic regression

1. Introduction

With an increasing amount of scientific medical literature being published each year, it is almost impossible to keep abreast of the actual status of knowledge in any specific field [1]. Summarizing the evidence by systematically reviewing the available literature and pooling estimates in a meta-analysis, when possible and useful, is also undergoing rapid expansion [2]. In order to summarize the relevant evidence, the articles at issue have to be identified. The initial sources used in the identification of relevant publications are often bibliographical databases, such as Medline [3]. Several authors have stressed the importance of an accurate search strategy that incorporates both high sensitivity and high specificity [4,5] for identifying relevant publications.

Search strategies for literature on clinical trials are available [3,6–8], but search strategies for literature on the evaluation of diagnostic tests have been less well studied [9,10]. The strategies that have been published were developed for diagnostic publications, including both publications specifically evaluating diagnostic tests and publications on clinical diagnosis not focusing on the evaluation of specific diagnostic signs, symptoms, or tests [7,9]. Other search strategies have been developed for the evaluation of a specific diagnostic test [11]. Overall, sensitive searches tend to be weak in specificity, resulting in the selection of a large number of irrelevant publications. In view of this high number of false positives, we wondered if it would be possible to develop a more specific search strategy for selecting publications on diagnostic test evaluations without loosing sensitivity, compared to the earlier diagnostic search strategies. We were not interested in any specific field, but in a standard strategy that could be used in any field of interest, as is the case for the strategies developed for identifying clinical trials [3].

* Corresponding author. Tel.: +31-20-444 8166; fax: +31-20-444 8181.
E-mail address: w.deville.emgo@med.vu.nl (W.L.J.M. Devillé)

0895-4356/00/$ – see front matter
PII: S0895-4356(99)00144-4

2. Methods

2.1. Selection of a reference set of publications by hand search

In 1990 the Departments of Family Medicine of the eight Medical Faculties in the Netherlands issued a ranking of the most relevant medical journals in their field. For our study, the 12 highest ranking journals were selected. Three journals—*Family Systems Medicine, Huisarts en Wetenschap* (Dutch), and *Huisarts Nu* (Belgian)—were not available in Medline, so the search was limited to the nine journals presented in Table 1. A search was started, covering publications from 1992 through 1995.

We intended to develop a search strategy for primary publications on diagnostic test evaluation. Primary publications excluded editorials, comments, news, reviews, and meta-analyses. We also excluded case reports and publications on animal research. Of these primary publications, we only included publications in which at least one diagnostic "test" (including clinical information) was compared with a "reference standard."

The nine journals were hand searched by WD (MD), blinded for any Medline search results. Titles and abstracts were screened. Publications with titles and/or abstracts related to case reports or treatment were skipped. If title and/or abstract gave any indication of diagnostic content (diagnostic keywords or text words, name of tests) or if there were any doubts about the content, the entire publication was read. All primary publications on diagnostic test evaluation were registered. This hand search was denoted the reference standard.

The publications detected by the reference standard were looked up in Medline and all Mesh terms (Medline subheading terms) and text words related to the field of diagnosis or test evaluation were noted for each publication, as presented in Medline.

Table 1
Primary papers on evaluations of diagnostic tests in 9 family medicine journals, 1992–1995

Journals	Papers	Diagnostic test (%)
American Family Physician	1299	0 (–)
British Journal of General Practice	786	5 (0.6)
Canadian Family Physician	635	2 (0.3)
Family Medicine	476	6 (1.2)
Family Practice	289	12 (4.2)
Family Practice Research Journal	134	3 (2.2)
Journal of Family Practice	996	28 (2.8)
Practitioner	498	0 (–)
Scandinavian Journal of Primary Health Care	204	15 (7.4)
All papers	5088	75 (1.5)
Primary papers[a]	2467	75 (3.0)

[a]Excluding editorials, comments, news, reviews, meta-analyses, case reports, and animal research.

2.2. Selection of a "control set" of publications

To develop and test a model for an optimal search strategy, we needed a set of publications that did not concern diagnostic test evaluation. To overcome the burden to register all possible Mesh terms from all 2467 primary publications, we looked for an alternative. Using methods similar to a nested case-control design, we decided to use the false-positive papers selected by a previously published diagnostic search strategy as a "control" set.

As, at that moment, only Haynes [7] had published an extensive paper on search strategies for diagnostic publications, his most sensitive and most specific search strategies for a diagnostic publications in 1991 were used in combination. The most sensitive search (sensitivity 0.92) combined the Mesh terms "sensitivity and specificity" exploded (exploded means also including "predictive value" and "ROC"), "diagnosis" (all subheadings, but not exploded), and "diagnostic use," together with the text words "sensitivity" and "specificity." The most specific search (specificity 0.98) combined only the exploded Mesh term "sensitivity and specificity" and the text words "predictive value" (or values). So, we added the text words "predictive value" to his most sensitive search strategy.

To limit the results of the above-mentioned searches to the nine journals included in our study, we selected these journals in Medline and combined their references by means of the Boolean term OR (terms enabling combinations of search terms), which produced the entire set of references published in these journals and available in Medline. This set of references was merged with the results of the Haynes searches by means of the Boolean AND, resulting in a set of diagnostic references for these journals.

Subsequently, the resulting set of references was first limited to the years 1992 through 1995, and secondly to primary publications, by excluding reviews, meta-analyses, comments, editorials, and news in the limitation feature under "JOURNAL TYPE." It was further limited by excluding case reports (Mesh term) and papers on animal research.

2.3. Development of the new search strategy

Univariate analysis was used to calculate sensitivity, specificity, and the diagnostic odds ratio of all relevant Mesh terms and diagnostic text words. Sensitivity of a search term was defined as the proportion of the publications of the reference set identified by the search term, specificity as the proportion of the publications in the control set correctly classified as controls, and the diagnostic odds ratio (DOR = positive likelihood ratio/negative likehood ratio) as a parameter for discrimination between the reference set of publications and the control set.

Models for a search strategy were developed by forward stepwise logistic regression analysis aiming at a correct classification of these publications into the categories "test evaluation" or "non-test evaluation."

All diagnostic Mesh terms and text words were options

for selection in the model in two phases. Firstly, only Mesh terms were options for inclusion and, secondly, in the resulting model the text words were added as options. Finally, the program presented the most predictive combinations of Mesh terms and text words.

The models were validated on the same sample of journals, by using them as search strategies in Medline and by comparing them with the findings of the most sensitive search strategy proposed by Haynes.

Sensitivity, specificity (with confidence intervals), and diagnostic odds ratios were calculated for four cumulative models with increasing numbers of text words. Sensitivity of a model was similar to the definition used for individual search terms. Specificity of a model was defined as the correctly classified proportion of all primary publications not belonging to the reference set. Diagnostic odds ratios were calculated as parameters for the power of the model to discriminate between the reference set and the other publications. We also calculated the positive predictive value for the complete number of primary publications, defined as the proportion of reference publications among all publications selected by the search ("precision" used in other bibliographic publications).

To test the generalizibility of the search strategy to a specific topic in the complete Medline database, we used our most extended search strategy for a review on the accuracy of physical diagnostic tests for the diagnosis of meniscal lesions of the knee. The results were compared with those of the Haynes strategy on the same topic.

3. Results

Hand searching identified 75 (3%) primary publications on test evaluation from a total of 2467 papers in the nine journals from 1992 through 1995 (these 75 publications were denoted the reference set) (Table 1).

The sensitive Haynes search found 192 publications on diagnosis, of which 55 concerned diagnostic test evaluation according to the reference standard; 137 publications were therefore classified as false positive according to the reference standard, and formed the control set. The publications were cross-checked in the journals concerned, and none corresponded with our definition of primary publications on diagnostic test evaluation.

Table 2 shows the list of Mesh terms and text words related to diagnosis or test evaluation with the respective number of true positives, sensitivity, number of false positives, specificity, and diagnostic odds ratio for discriminating the reference set from the control set.

Table 2
Univariate estimates of Mesh terms and text words to discriminate publications on diagnostic test evaluation ($n = 75$) from diagnostic publications not concerning test evaluation ($n = 137$)

	True positives (sensitivity %)	False positives (specificity %)	DOR
Mesh term			
Sensitivity and specificity	38 (51)	16 (88)	7.6
Diagnosis[a]	51 (68)	66 (52)	2.3
Predicitive value	18 (24)	13 (91)	3.2
ROC	6 (8)	2 (99)	8.6
Mass screening	11 (15)	9 (93)	2.3
Reproducibility	10 (13)	6 (96)	3.6
False-positive reactions	3 (4)	1 (99)	4.1
False-negative reactions	5 (7)	2 (99)	7.5
Logistic modeling	2 (3)	1 (99)	3.1
Regression analysis	4 (5)	0 (100)	—
Sensitivity and specificity (exploded with predictive value and ROC)	47 (63)	29 (79)	6.4
Text word[b]			
Sensitivity	31 (41)	20 (85)	3.9
Specificity	26 (35)	5 (96)	12.9
Diagnos:[c]	35 (47)	96 (30)	0.38
Predictive value	14 (19)	6 (96)	5.6
ROC	4 (5)	1 (99)	5.2
Screening	24 (32)	17 (88)	3.5
Reproducibility (or reliability)	7 (9)	5 (96)	2.4
False positive	3 (4)	0 (100)	—
False negative	6 (8)	1 (99)	8.6
Logistic regression	7 (9)	2 (99)	9.8
Likelihood ratio	4 (5)	2 (99)	5.2
Accuracy	10 (13)	5 (96)	3.6

DOR = diagnostic OR.
[a]Mesh using "diagnosis."
[b]Text words in title or abstract.
[c]Text words starting with "diagnos."

The Mesh terms "sensitivity and specificity" (which together form one Mesh term) and "diagnosis" proved to be the most sensitive ones. On the other hand, "diagnosis" had the lowest specificity, resulting in the lowest diagnostic odds ratio, together with "mass screening." The Mesh term "sensitivity and specificity" had a three times higher discriminative power than "diagnosis." "False-negative reactions" and "ROC" also had relatively high diagnostic odds ratios, but low sensitivities.

For the text words, sensitivities were lower than for the corresponding Mesh terms, but the words "sensitivity" and "specificity," and all forms of the word "diagnosis," had higher sensitivities than the other text words, as was the case for the Mesh terms, except for "screening." Of these four text words "specificity" had the highest specificity in the detection of the relevant publications, while the text word "diagnosis" had a very high number of false positives, resulting in a diagnostic odds ratio of less than 1. From these only "specificity" had a very high diagnostic odds ratio.

Logistic modeling, aimed at optimal discrimination between the two sets of publications, resulted in the following.

With only Mesh terms, the model ended up with the terms "sensitivity and specificity" (exploded: including "predictive value" and "ROC"). It correctly classified 47 of the 75 (63%) publications on diagnostic test evaluation and 108 (795) of the 137 publications in the control set.

Expanding the above model with text words resulted in a set of five terms: the Mesh term "sensitivity and specificity" (exploded) and, cumulatively, four text words: "specificity," "false negative," "accuracy," and "screening." The optimal model here included only "sensitivity and specificity" (exploded) and the text word "specificity": it correctly classified 54 (72%) of the reference set and 105 (77%) of the control set.

The four different models including text words were subsequently used as search strategies in Medline for primary publications on diagnostic test evaluation in the nine journals from 1992 through 1995, and compared with the results of the search strategy used by Haynes. This resulted in the identification of at least 53 (71%) and at most 67 (89%) of the 75 papers on diagnostic test evaluation, compared to the 55 identified by the sensitive Haynes strategy.

On the other hand, 36 to 193 false-positive publications were selected compared to the 137 selected by the Haynes search. Adding the text word "screening" to the strategy increased the sensitivity from 80% to 89%, but decreased the specificity from 97% to 92%, resulting in twice the number of false-positive abstracts.

The most accurate Medline search strategy combined the Mesh term "sensitivity and specificity" (exploded) with the text words "specificity," "false negative," and "accuracy." It resulted in a sensitivity of 80.0% (60/75; 95% CI: 71.0–89.1) and a specificity of 97.3% (2327/2392; 95% CI: 96.6–97.9), both higher than the sensitive Haynes search with 73.3% (95% CI: 63.3–83.3) and 94.3% (95% CI: 93.3–95.2), respectively. Positive predictive value and DOR were 48% (95%CI: 40–56) and 143 for our strategy, versus 29% (95%CI: 23–35) and 45 for the Haynes search, respectively. The various strategies and their validity estimates can be found in Table 3.

The most sensitive search strategy (No. 4, Table 3) was used in another standard set including 33 papers on physical diagnostic tests for meniscal lesions. Our strategy resulted in a sensitivity of 61% (20/33) and a predictive value of

Table 3
Validity estimates of search strategies for publications on diagnostic test evaluations in Medline

Search strategy	Content	Sensitivity % (95% CI) ($N = 75$)	Specificity % (95% CI) ($N = 2392$)	DOR
Haynes' sensitive strategy	sensitivity and specificity (exploded) (sh) diagnosis& (sh) diagnostic use (sh) sensitivity (tw) specificity (tw)	73.3 (63.3–83.3)	94.3 (93.3–95.2)	45
Strategy 1	sensitivity and specificity (exploded) (sh) specificity (tw)	70.7 (60.4–81.0)	98.5 (98.0–98.9)	158
Strategy 2	sensitivity and specificity (exploded) (sh) specificity (tw) false negative (tw)	73.3 (63.3–83.3)	98.4 (97.9–98.9)	170
Strategy 3	sensitivity and specificity (exploded) (sh) specificity (tw) false negative (tw) accuracy (tw)	80.0 (71.0–89.1)	97.3 (96.6–97.9)	143
Strategy 4	sensitivity and specificity (exploded) (sh) specificity (tw) false negative (tw) accuracy (tw) screening (tw)	89.3 (82.3–96.3)	91.9 (90.8–93.0)	95

sh = Medline subheading; tw = text word.

4.7% (20/428), versus 45% (15/33) and 3.4% (15/441) for the most sensitive Haynes strategy.

4. Discussion

When using computerized bibliographical databases for literature searches we are confronted with various possibilities and limitations. Advantages of computerized bibliographical databases are that they cover a considerable amount of the most important medical literature and are efficient to use if one has some experience. However, they cover only a limited number of journals, encounter various delays in entering publications for different journals, and the labeling according to keywords—as Mesh terms in Medline—is not completely accurate. On the other hand, searching by hand is cumbersome and time consuming [5].

A well-considered and accurate search strategy is, of course, of the utmost importance. The new model including the text words "specificity," "false negative," and "accuracy," in addition to the Mesh term "sensitivity and specificity" (exploded), is more sensitive, and its specificity is significantly higher than the most sensitive Haynes search strategy for diagnostic publications [7]. The increase in specificity is small, but in our example it reduces by half the absolute number of false-positive abstracts that have to be screened. The strategy (including only the text words "specificity" and "false negative" in addition to the Mesh term), with an identical sensitivity to that of Haynes, reduced the number of false positives to a quarter. The most sensitive strategy, adding the text word "screening," decreased specificity considerably. This strategy identifies publications on the evaluation of screening activities that have not been detected by other strategies.

Haynes'search strategies were developed on a set of publications of 10 major peer-reviewed journals of internal medicine and general medicine from the years 1986 and 1991. Inclusion of reviews and case reports, as well as the definition used for a diagnostic publication—"Content pertained directly to the evaluation of a disease process" [7]—may be an explanation of the higher accuracy obtained in his study.

While the need for the highest accuracy—the highest possible sensitivity combined with the highest possible specificity—is obvious, the balance between sensitivity and specificity can be a point of discussion. In principle, the intention is to identify all publications in the field of interest, but the most sensitive search is weak in specificity. This will not be a problem in fields of research with a limited number of publications, but in fields with a considerable number of publications, it can imply that a great number of abstracts have to be screened, which will later have to excluded. The purpose of the search is therefore important. If publications are being collected for a systematic review or a meta-analysis, one has to accept search strategies with a low specificity. Future research should evaluate in which way a more accurate, but less sensitive, search strategy affects the conclusions or pooled estimates in reviews and meta-analyses, as the most important studies will probably be detected anyway.

When reviewing the literature to obtain initial evidence for the development of research protocols, discussing the literature with residents or students, or preparing lectures, less sensitive but more specific searches can be less time consuming, and thus more suitable. On the other hand, if the interest is directed towards publications on screening activities, it is better to use a strategy including the word "screening."

The newly developed strategies have the advantage that they are specifically focused on diagnostic literature concerning test evaluation (i.e., the literature needed to perform a diagnostic meta-analysis). They can be applied to any field of interest, as is the case of the strategies developed for clinical trials [3]. Whether they really have an advantage over the recently published subject-specific search strategies [11], once subject-specific search terms are added, has yet to be established and, if possible, should be tested on the same data. The results obtained for the field of meniscal lesions of the knee seem promising, but also show the still limited sensitivity of these strategies in different areas of medical research. The newly designed strategies should be validated on other sections of the medical literature.

Acknowledgment

We thank Ms. D. van der Windt for critically reviewing the manuscript.

References

[1] Buntinx F. The Cochrane collaboration, information overload and European general practice. Eur J Gen Pract 1995;1:11–12.
[2] Mulrow CD. Rationale for systematic reviews. BMJ 1994;309:597–9.
[3] Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309:1286–91.
[4] Lowe HJ, Barnett GO. Understanding and using the Medical subject Headings (Mesh) vocabulary to perform literature searches. JAMA 1994;271:1103–8.
[5] Jadad AR, McQuay HJ. Be systematic in your searching. BMJ 1993; 307:66.
[6] Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. Online J Curr Clin Trials 1993; Doc No 33.
[7] Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in Medline. J Am Med Informatics Assoc 1994;1:447–58.
[8] Marson AG, Chadwick DW. How easy are randomized controlled trials in epilepsy to find on Medline? The sensitivity and precision of two Medline searches. Epilepsia 1996; 37:377–80.
[9] McKibbon KA. Beyond the ACP Journal Club: how to harness Medline for diagnostic problems. ACP Journal Club 1994 Sept/Oct A10–12.
[10] Irwig L, Tosteson ANA, Gastonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluation diagnostic test. Ann Intern Med 1994;120:667–76.
[11] van der Weijden T, Ijzermans CJ, Dinant G, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in Medline. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. Fam Prac 1997;14:204–8.